Homework 1

Mark Schulist

1.

a. Each row represents an observational unit in the data set, in this case, they represent UK residents.

b. There were 1691 participants in the study—it's the number of rows in the data set.

c. sex: nominal; age: discrete; marital: nomial; grossIncome: ordinal; smoke: nominal; amtWeekends: ordinal; amtWeekdays: ordinal

2.

a. No, this is an observational study and not an experiment. They did not prescribe treatments to different randomized groups.

b. Maybe drinking more coffee makes people have more cramps. Less sleep could also make people get cramps. These are counfounding variables that might indirectly be the cause of the relationship between stress and muscle cramps.

3.

a.

b. Depending on how you split the neighborhoods, you might over represent certain neighborhoods.

c. You will miss out on $\frac{17}{20}$ of the neighborhoods, but you will get a good sample from the ones you sample. This strategy could be good if the neighborhoods you sample are representative of the rest of the city.

9.

```r
library(tidyverse)
library(gridExtra)
source("https://www.openintro.org/data/R/acs12.R")

median = median(acs12$income, na.rm = T)
```

```
IQR = IQR(acs12$income, na.rm = T)

grad_prop <- table(acs12$edu)["grad"] / length(acs12$edu)

acs12 <- acs12 %>%
  mutate(
    edu_int = as.numeric(edu)
  )

income <- ggplot(acs12) +
  geom_boxplot(aes(income))
edu <- ggplot(acs12) +
  geom_boxplot(aes(edu_int))

grid.arrange(income, edu)
```
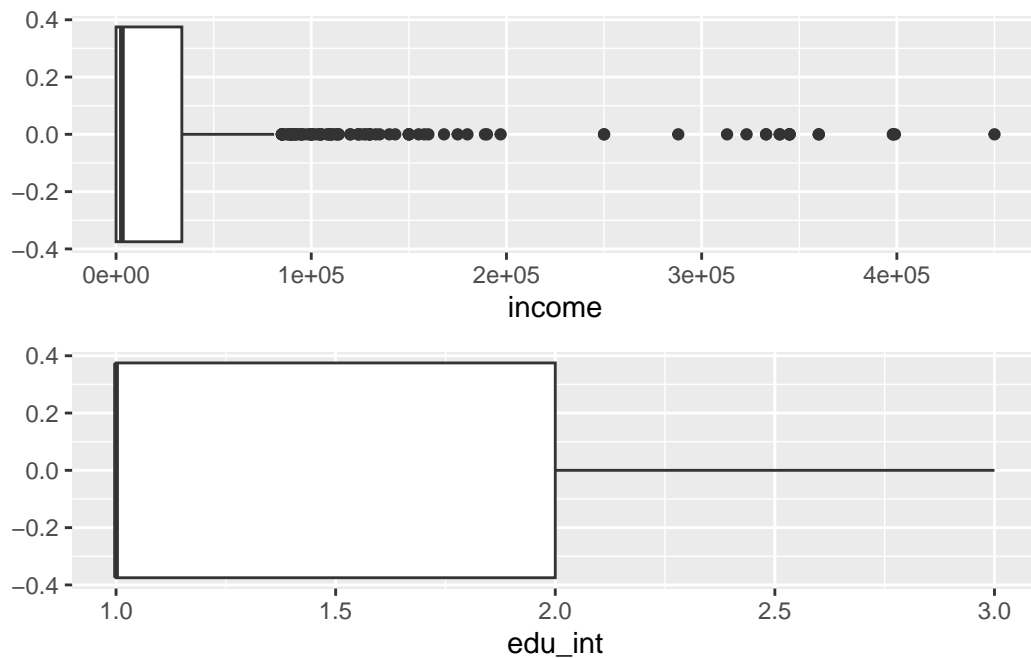

Median $= 3000$

IQR $= 3.37 \times 10^4$

Grad Proportion $= 0.072$